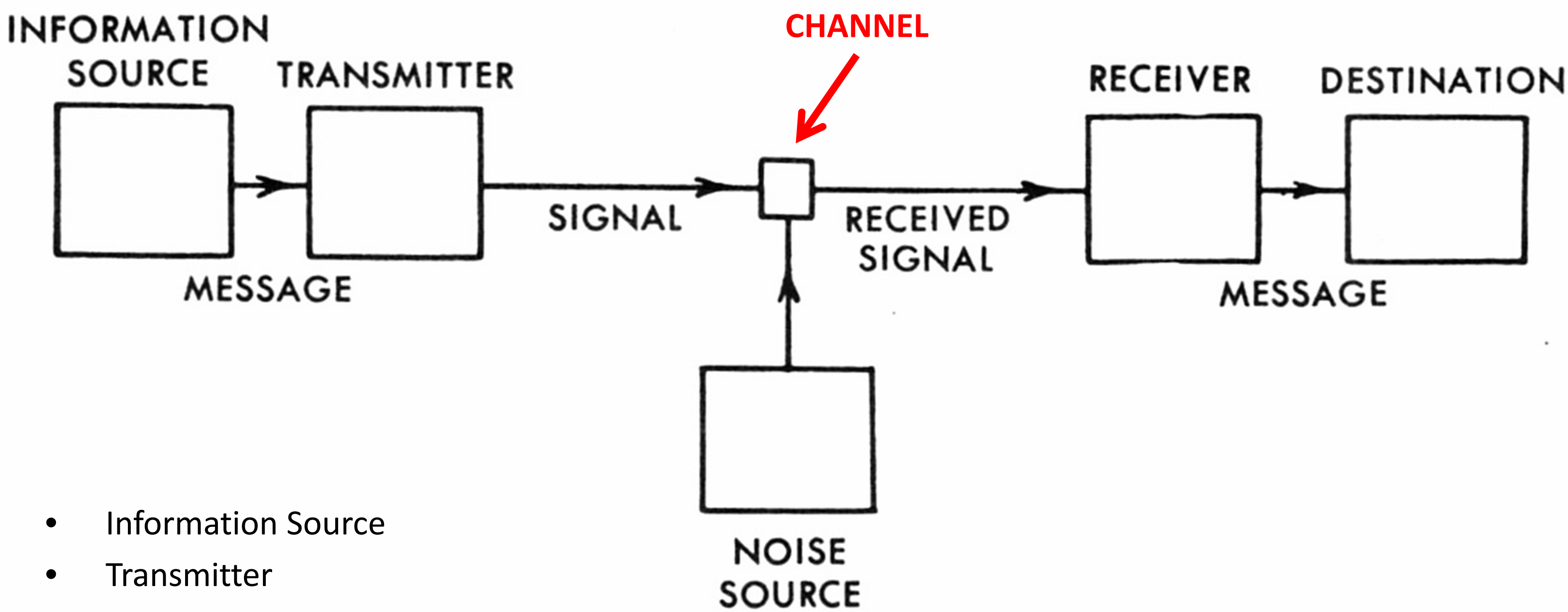
Introduction to Information Theory

Part 2

A General Communication System



- Information Source
- Transmitter
- Channel
- Receiver
- Destination

Information: Definition

- Information is quantified using probabilities.
- Given a finite set of possible messages, associate a probability with each message.
- A message with low probability represents more information than one with high probability.

Definition of Information:

$$\boldsymbol{I(p) = log\left(\frac{1}{p}\right) = -log(p)}$$

Where $\boldsymbol{p}$ is the probability of the message

Base 2 is used for the logarithm so $\boldsymbol{I}$ is measured in **bits**

Trits for base 3, **nats** for base e, **Hartleys** for base 10…

$$\boldsymbol{I(p) = \log\left(1/p\right) = -\log(p)}$$

Some properties of I

1. $I(p) \geq 0$

   Information is non-negative.

2. $I(p_1 * p_2) = I(p_1) + I(p_1)$

   Information we get from observing two independent events occurring is the sum of two information(s).

3. $I(p)$ is monotonic and continuous in p

   Slight changes in probability incur slight changes in information.

4. $I(p = 1) = 0$

   We get zero information from an event whose probability is 1.

Example: Information in a coin flip

$$p_{HEADS} = 1/2$$

$$I_{HEADS} = -\log(1/2) = 1 bit$$

Independent Events: 2 Coin flips

- There are four possibilities: HH, HT, TH, TT

$$I_{HH} = \log\left(\frac{1}{p_H * p_H}\right) = \log\left(\frac{1}{1/4}\right) = \log(4) = 2$$

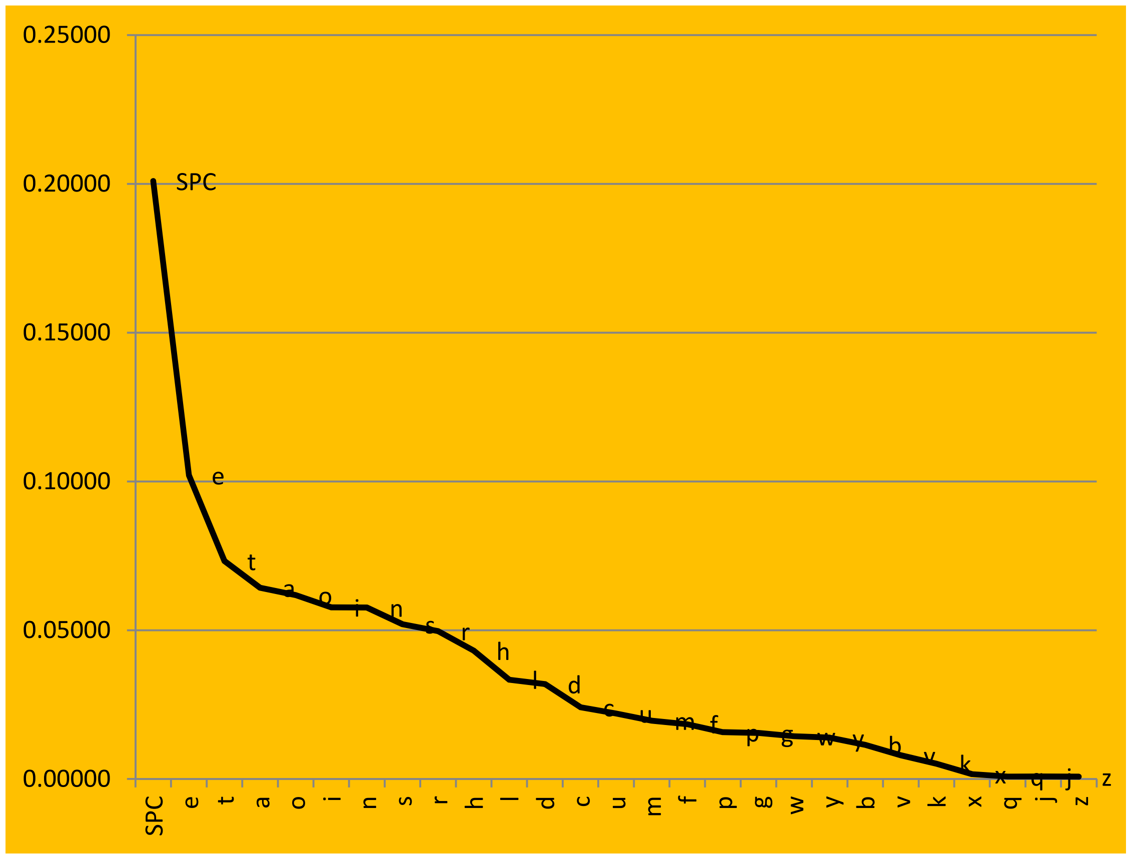
i.e. Additive property:

$$I_{AB} = -\log(p_A p_B) = -\log(p_A) - \log(p_B)$$

$$I_{AB} = I_A + I_B$$

Example: Text Analysis

a	0.06428
b	0.01147
c	0.02413
d	0.03188
e	0.10210
f	0.01842
g	0.01543
h	0.04313
i	0.05767
j	0.00082
k	0.00514
l	0.03338
m	0.01959
n	0.05761
o	0.06179
p	0.01571
q	0.00084
r	0.04973
s	0.05199
t	0.07327
u	0.02201
v	0.00800
w	0.01439
x	0.00162
y	0.01387
z	0.00077
SPC	0.20096



Example: Text Analysis

Letter	Freq.	I
a	0.06428	3.95951
b	0.01147	6.44597
c	0.02413	5.37297
d	0.03188	4.97116
e	0.10210	3.29188
f	0.01842	5.76293
g	0.01543	6.01840
h	0.04313	4.53514
i	0.05767	4.11611
j	0.00082	10.24909
k	0.00514	7.60474
l	0.03338	4.90474
m	0.01959	5.67385
n	0.05761	4.11743
o	0.06179	4.01654
p	0.01571	5.99226
q	0.00084	10.21486
r	0.04973	4.32981
s	0.05199	4.26552
t	0.07327	3.77056
u	0.02201	5.50592
v	0.00800	6.96640
w	0.01439	6.11899
x	0.00162	9.26697
y	0.01387	6.17152
z	0.00077	10.34877
SPC	0.20096	2.31502

Definition of Entropy

- Information (I) is associated with known events/messages
- Entropy (H) is the average information w.r.to all possible outcomes.

Given, $P = \{p_1, p_2, \dots, p_3\}$

$$H(P) = \sum_i p_i \log(\frac{1}{p_i})$$

Characterizes an **information source**.

Example: A 3-event Source

$$A = \{a_1, a_2, a_3)$$

$$P = \{p, p_2, p_3\} = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\}$$

$$H(P) = \frac{1}{2}\log(2) + \frac{1}{4}\log(4) + \frac{1}{4}\log(4)$$

$$= \frac{1}{2} + \frac{1}{4} * 2 + \frac{1}{4} * 2 = \frac{3}{2} = 1.5\ bits$$

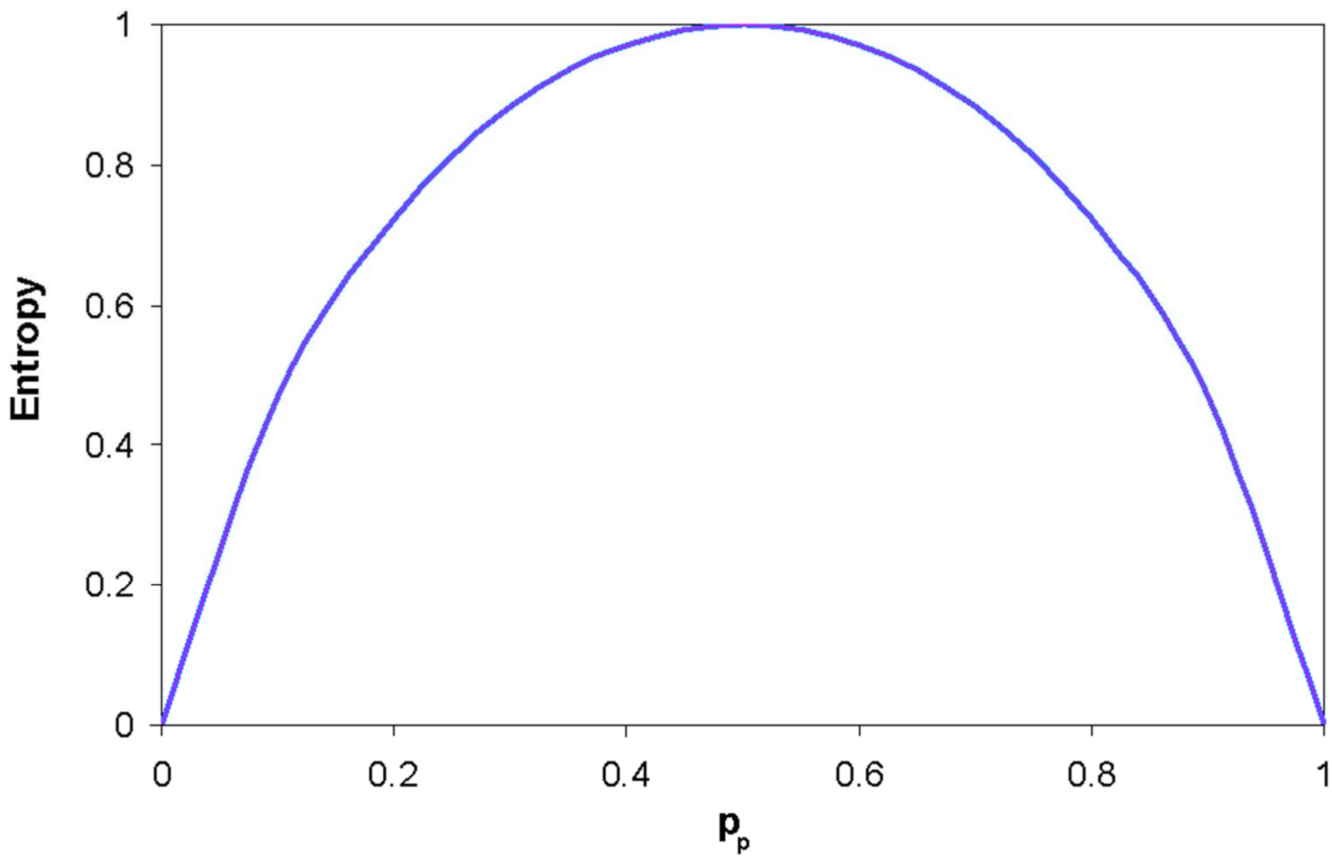
Example: Text Analysis

Letter	Freq.	I
a	0.06428	3.95951
b	0.01147	6.44597
c	0.02413	5.37297
d	0.03188	4.97116
e	0.10210	3.29188
f	0.01842	5.76293
g	0.01543	6.01840
h	0.04313	4.53514
i	0.05767	4.11611
j	0.00082	10.24909
k	0.00514	7.60474
l	0.03338	4.90474
m	0.01959	5.67385
n	0.05761	4.11743
o	0.06179	4.01654
p	0.01571	5.99226
q	0.00084	10.21486
r	0.04973	4.32981
s	0.05199	4.26552
t	0.07327	3.77056
u	0.02201	5.50592
v	0.00800	6.96640
w	0.01439	6.11899
x	0.00162	9.26697
y	0.01387	6.17152
z	0.00077	10.34877
SPC	0.20096	2.31502

$$H(P) = \sum_i p_i \log\left(\frac{1}{p_i}\right) = 4.047$$

Aka, First-Order Entropy.

Entropy (2 outcomes)



Entropy: Properties

1. $H(P) \geq 0$

2. $H(P) \leq \log(n)$
   Entropy is maximized if P is uniform.

3. $H(S, T) = H(S) + H(T)$
   Additive property for independent events.

4. $H(S, T) \leq H(S) + H(T)$
   If S and T are not independent.

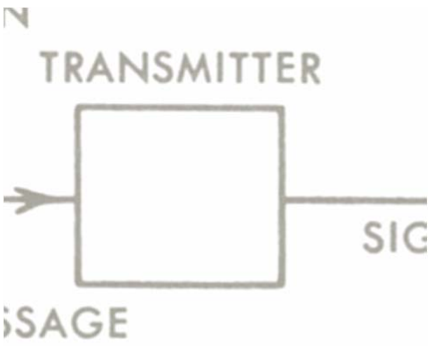
Entropy of things…

- Entropy of English text is approx 1.5 bits
- Entropy of the human genome <= 2 bits
- Entropy of a black hole is ¼ of the area of the outer event horizon.
- Value of information in economics is defined in terms of entropy. E.g. Scarcity

$$V(X) = \sum_{i=1}^{n} p_i(-\log_b(p_i))$$

Entropy: What about it?

- Does $H(P)$ have a maximum? Where?
- Is entropy a good name for this stuff? How is it related to entropy in thermodynamics?
- How does entropy help in communication? What else can we do with it?
- Why use the letter H? ☺



Entropy & Codes

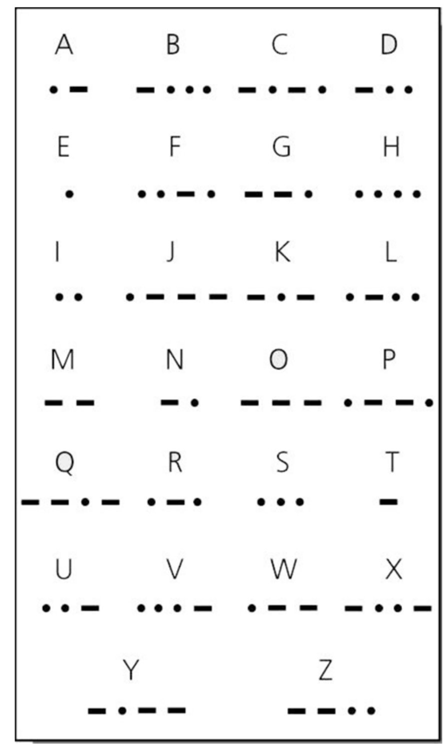
- Entropy is closely related to the design of efficient codes for random sources.
- Provides foundations for techniques of compression, data search, encryption, correction of communication errors, etc.
- Essential to the study of life sciences, economics, etc.

Coding: Basics

- Events of an information source: $s_1, s_2, \ldots, s_m$

- A **code** is made up of **codewords** from a **code alphabet**
(e.g. {0, 1}, {., -}, etc.)

- A **code** is an assignment of codewords to source symbols.



Coding: Basics

- **Block code**: When all codes have the same length. For example, ASCII (8-bits)

- **Average Word Length**:

$$L = \sum_{i=1}^{m} p_i l_i$$

More generally,

$$L_n = \frac{1}{n} \sum_{i=1}^{m} p_i l_i$$

- A code is **efficient** if it has the smallest average word length. (Turns out entropy is the benchmark…)

Coding: Basics

- **Singular** (not unique) codes
- **Nonsingular** (unique) codes

Symbol	Singular Code	Nonsingular Code
A	00	0
B	10	10
C	01	00
D	10	01

Coding: Basics

- **Singular** (not unique) codes
- **Nonsingular** (unique) codes
- **instantaneous** codes
  (every word can be decoded as soon as it is received)

Symbol	Singular Code	Nonsingular Code
A	00	0
B	10	10
C	01	00
D	10	01

Not an instantaneous Code!

Example: Avg. Code Length (L)

Symbol	p	Codeword
A	0.3	00
B	0.2	10
C	0.2	11
D	0.2	010
E	0.1	011

$$L = (0.3 * 2) + (0.2 * 2) + (0.2 * 2) + (0.2 * 3) + (0.1 * 3) = 2.3$$

Example: Source Entropy (H)

Symbol	p	Codeword
A	0.3	00
B	0.2	10
C	0.2	11
D	0.2	010
E	0.1	011

$$L = (0.3 * 2) + (0.2 * 2) + (0.2 * 2) + (0.2 * 3) + (0.1 * 3) = 2.3$$

$$H = 0.3 \log\left(\frac{1}{0.3}\right) + 0.2 \log\left(\frac{1}{0.2}\right) * 3 + 0.1 \log\left(\frac{1}{0.1}\right) = 2.246$$

Example: L & H

Symbol	p	Codeword
A	0.3	00
B	0.2	10
C	0.2	11
D	0.2	010
E	0.1	011

$$L = (0.3 * 2) + (0.2 * 2) + (0.2 * 2) + (0.2 * 3) + (0.1 * 3) = 2.3$$

$$H = 0.3 \log\left(\frac{1}{0.3}\right) + 0.2 \log\left(\frac{1}{0.2}\right) * 3 + 0.1 \log\left(\frac{1}{0.1}\right) = 2.246$$

Is there a relationship between L and H?

Average Code Length & Entropy

- Average length bounds: $H \leq L < H + 1$

- Grouping n symbols together:

$$H(S^n) \leq L < H(S^n) + 1$$

Average Code Length & Entropy

- Average length bounds: $H \leq L < H + 1$

- Grouping n symbols together:

$$H(S^n) \leq L < H(S^n) + 1$$

$$nH(S) \leq L < nH(S) + 1$$

Average Code Length & Entropy

- Average length bounds: $H \leq L < H + 1$
- Grouping n symbols together:

$$H(S^n) \leq L < H(S^n) + 1$$

$$nH(S) \leq L < nH(S) + 1$$

$$H(S) \leq \frac{L}{n} < H(S) + \frac{1}{n}$$

Average Code Length & Entropy

- Average length bounds: $H \leq L < H + 1$

- Grouping n symbols together:

$$H(S^n) \leq L < H(S^n) + 1$$

$$nH(S) \leq L < nH(S) + 1$$

$$H(S) \leq \frac{L}{n} < H(S) + \frac{1}{n}$$

Average Code Length & Entropy

- Average length bounds: $H \leq L < H + 1$
- Grouping n symbols together:

$$H(S^n) \leq L < H(S^n) + 1$$

$$nH(S) \leq L < nH(S) + 1$$

$$H(S) \leq \frac{L}{n} < H(S) + \frac{1}{n}$$

$$\lim_{n \to \infty} \frac{L_n}{n} = \mathrm{H}$$

Shannon's First Theorem

- By coding sequences of independent symbols (in S^n), it is possible to construct codes such that

$$\lim_{n \to \infty} \frac{L_n}{n} = \mathrm{H}$$

The price paid for such improvement is increased coding complexity (due to increased n) and increased delay in coding.

Entropy & Coding: Central Ideas

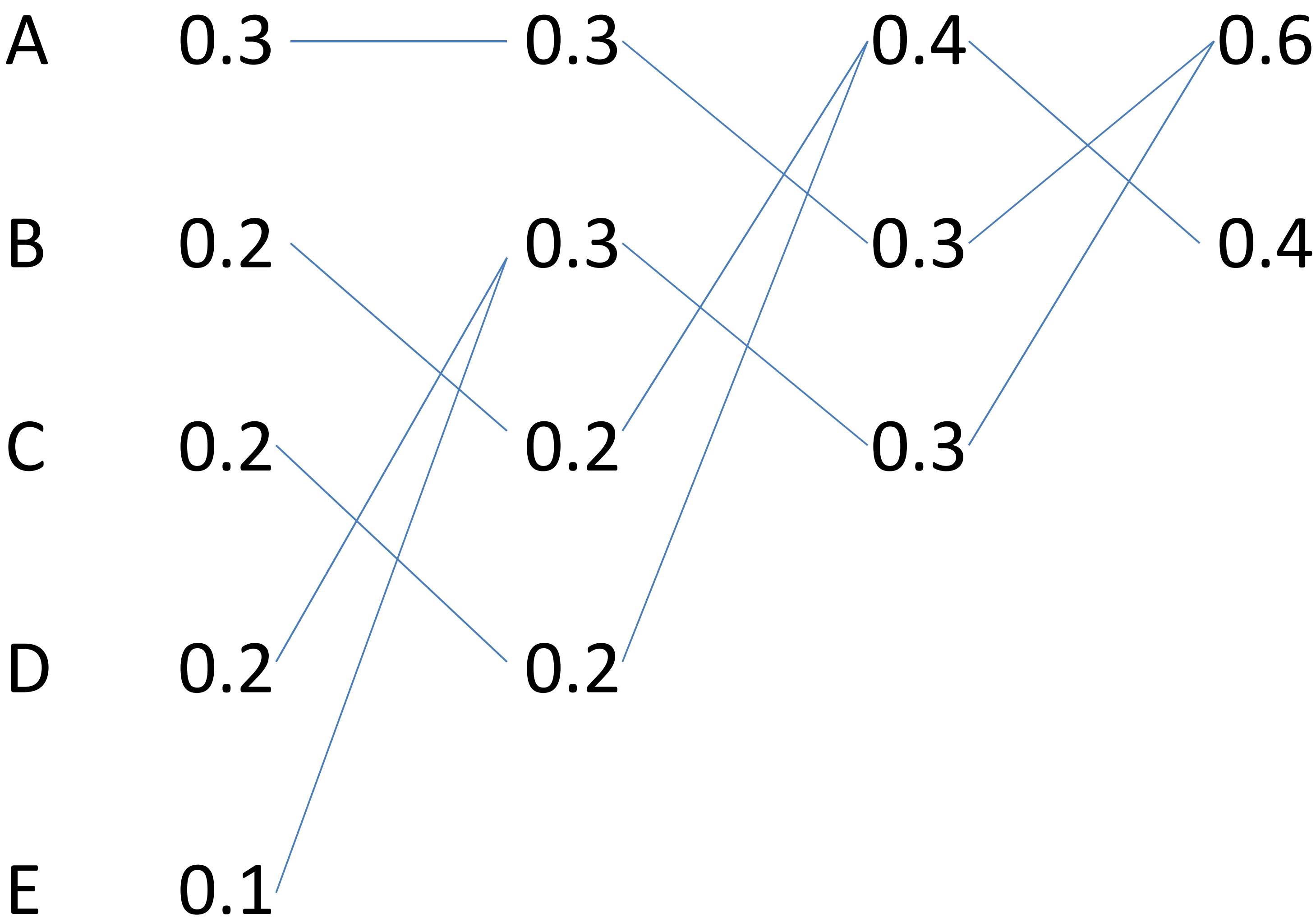
- Use short codes for highly likely events. This shortens the average length of coded messages.

- Code several events at a time. Provides greater flexibility in code design.

Data Compression: Huffman Coding

A	0.3
B	0.2
C	0.2
D	0.2
E	0.1

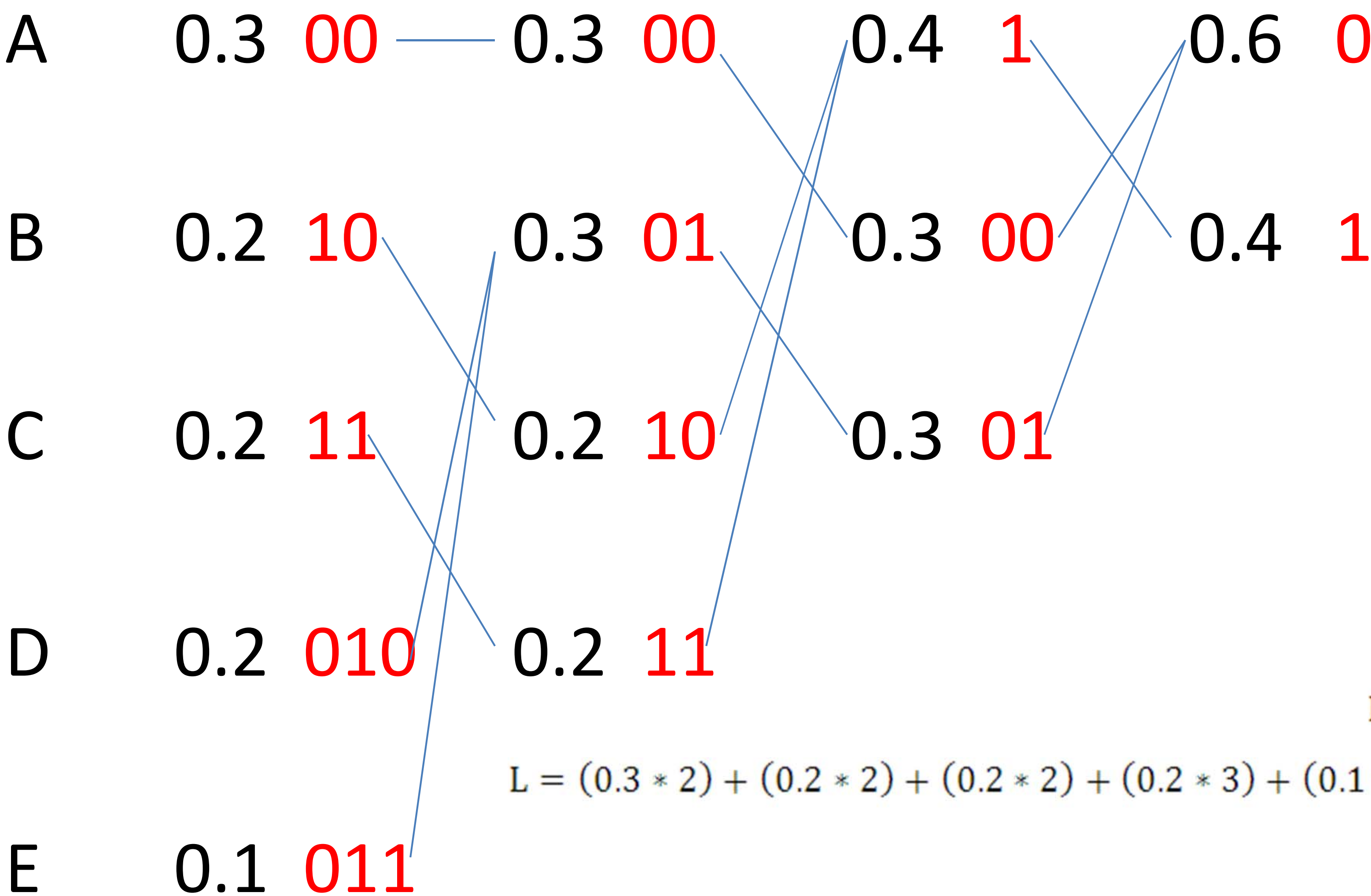
Huffman Coding: Reduction Phase



Huffman Coding: SplittingPhase

A	0.3	00	0.3	00	0.4	1	0.6	0
B	0.2	10	0.3	01	0.3	00	0.4	1
C	0.2	11	0.2	10	0.3	01		
D	0.2	010	0.2	11				
E	0.1	011						

Huffman Coding: SplittingPhase



$$H = 2.246$$

$$L = (0.3 * 2) + (0.2 * 2) + (0.2 * 2) + (0.2 * 3) + (0.1 * 3) = 2.3$$

Huffman Codes

- Nonsingular
- Instantaneous
- Efficient
- Non-unique
- Powers of a source lead closer to H
- Requires knowledge of symbol probabilities

Design Huffman Codes

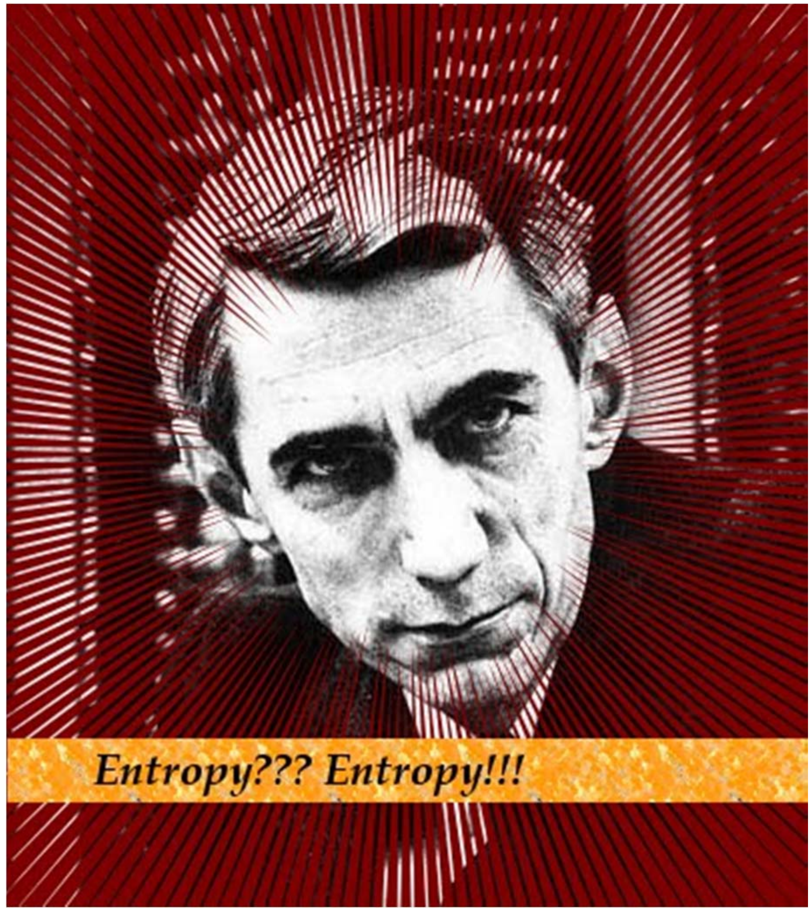
- $S = \{A, B\}, P = \{0.75, 0.25\}$

- $S = \{AA, AB, BA, BB\}$

- $S = \{AAA, AAB, ABA, BAA, ABB, BAB, BBA, BBB\}$

References

- Eugene Chiu, Jocelyn Lin, Brok Mcferron, Noshirwan Petigara, Satwiksai Seshasai: *Mathematical Theory of Claude Shannon: A study of the style and context of his work up to the genesis of information theory. MIT* 6.933J / STS.420J The Structure of Engineering Revolutions
- Luciano Floridi, 2010: *Information: A Very Short Introduction*, Oxford University Press, 2011.
- Luciano Floridi, 2011: *The Philosophy of Information*, Oxford University Press, 2011.
- James Gleick, 2011: *The Information: A History, A Theory, A Flood*, Pantheon Books, 2011.
- Zhandong Liu , Santosh S Venkatesh and Carlo C Maley, 2008: *Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples*, *BMC Genomics* 2008, **9:**509
- David Luenberger, 2006: *Information Science*, Princeton University Press, 2006.
- David J.C. MacKay, 2003: *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.
- Claude Shannon & Warren Weaver, 1949: *The Mathematical Theory of Communication*, University of Illinois Press, 1949.
- W. N. Francis and H. Kucera: *Brown University Standard Corpus of Present-Day American English*, Brown University, 1967.


Entropy??? Entropy!!!

$$\boldsymbol{H(S,T) = H(S) + H(T)}$$

Additive property.

S & T are independent sources,

$$H(S,T) = -\sum_{s\in S, t\in T} p_s p_t \log(p_s p_t)$$

$$= -\sum_{s\in S, t\in T} p_s p_t [\log(p_s) + \log(p_t)]$$

$$= -\sum_{t\in T} p_t \left[\sum_{s\in S} p_s \log(p_s)\right] - \sum_{s\in S} p_s \left[\sum_{t\in T} p_t \log(p_t)\right]$$

$$= H(S) + H(T)$$